Clustering

Unsupervised Learning

In unsupervised learning, there’s no labels or responses.

The goal is to find structure in data. Examples like clustering, dimensionality reduction, data compression and etc.

Clustering

Unsupervised learning, the process of partitioning a set of data into a set of meaningful sub-classes, called clusters.

Cluster:

- collection of data points similar to each other
- as a collection, are sufficiently different from other groups

Clustering methodology

Hierarchical Algorithms

- Agglomerative: pairs of items/clusters are successively linked to produce larger clusters
- Divisive (partitioning): items are initially placed in one cluster and then divided into separate groups

Flat Algorithms

Usually start with a random (partial) partitioning of points into groups, K-Means.

K-Means

Have a data set $\{x_1, \ldots, x_N\}$ consisting of N observations of a random D-dimensional Euclidean variable x, the goal is to partition the data set into some number K of clusters.

A cluster is a group of data points whose inter-point distances are small compared with the distances to points outside of the cluster.

A set of D-dimensional vectors μ_k, which μ_k is a prototype associated with the k_{th} cluster, and can *represent the centres of the clusters*.

Our goal is then to find an assignment of data points to clusters, as well as a set of vectors $\{\mu_k\}$, such that the sum of the squares of the distances of each data point to its closest vector μ_k, is a minimum.

Assignment of data points:

For each data point x_n, a corresponding set of binary indicator variables $r_{nk} \in \{0, 1\}$, describing which of the K clusters the data point x_n is assigned to, so that if data point x_n is assigned to cluster k then $r_{nk} = 1$, and $r_{nj} = 0$ for $j \neq k$ (1-of-K coding scheme) and define a function, called distortion measure

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \|x_n - \mu_k\|^2$$

which represents the sum of the squares of the distances of each data point to its assigned vector μ_k. Find values for $\{r_{nk}\}$ and $\{\mu_k\}$ so as to minimise J.

To do this, an iterative procedure in which each iteration involves 2 successive steps corresponding to successive optimisations with respect to the r_{nk} and the μ_k.

1. Choose some initial values for μ_k
2. 1st phase, minimise J with respect to r_{nk}, keep μ_k fixed
3. 2nd phase, minimise J with respect to μ_k, keep r_{nk} fixed
4. repeat until convergence.

> We shall see that these two stages of updating r_{nk} and updating μ_k correspond respectively to the E (expectation) and M (maximisation) steps of the EM algorithm, and to emphasize this we shall use the terms E step and M step in the context of the K-means algorithm.

Expectation step

Here J is a linear function of r_{nk}.

The terms involving different n are independent and so we can optimise for each n separately by choosing r_{nk} to be 1 for whichever value of k gives the minimum value of $\|x_n - \mu_k\|^2$.

Simply assign the n^{th} data point to the closest cluster centre.

$$r_{nk} = \begin{cases} 1, & \text{if } k = argmin_j \|x_n - \mu_j\|^2 \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

Maximisation step

Here J is a quadratic function of μ_k and it can be minimised by setting its derivative with respect to μ_k to zero

$$2 \sum_{n=1}^{N} r_{nk}(x_n - \mu_k) = 0$$

which can easily solve for μ_k to give

$$\mu_k = \frac{\sum_n r_{nk} x_n}{\sum_n r_{nk}}$$

The denominator in this expression is equal to the number of points assigned to cluster k, and so this result has a simple interpretation, namely set μ_k equal to the mean of all of the data points x_n assigned to cluster k. For this reason, the procedure is known as the K-means algorithm.

The 2 phases re-assigning the data points to clusters and re-computing the cluster means. The value of the objective function J is reducing during the repetition and converge to a local minimum instead of the global minimum of J.

Picking the initial values for μ_k

A better initialisation procedure would be to choose the cluster centres μ_k to be equal to a random subset of K data points.